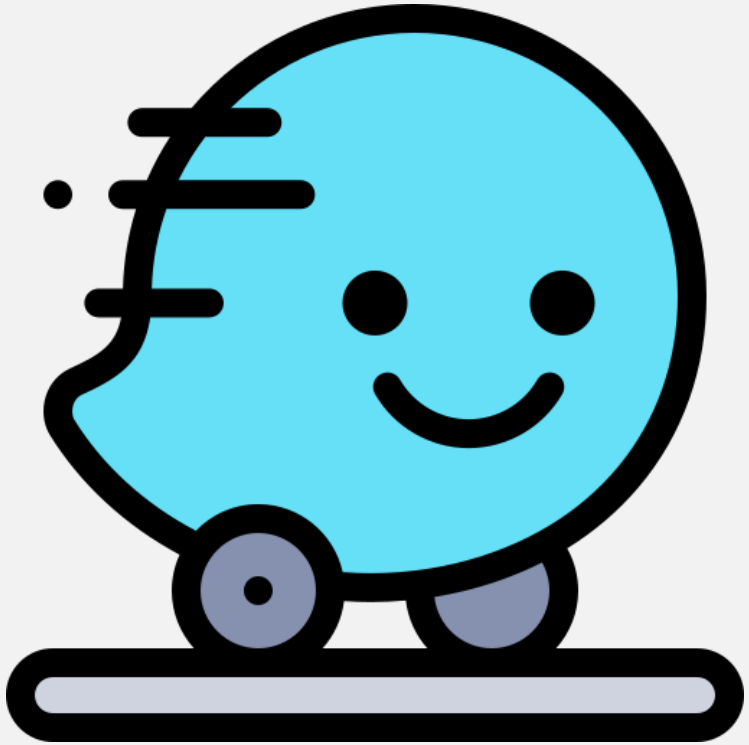
Mini Project Data Science



Project Report

Iswara Jay Jr.

Outline :

1. Background
2. Objectives
3. Data Understanding
4. Modelling
5. Results
6. Summary

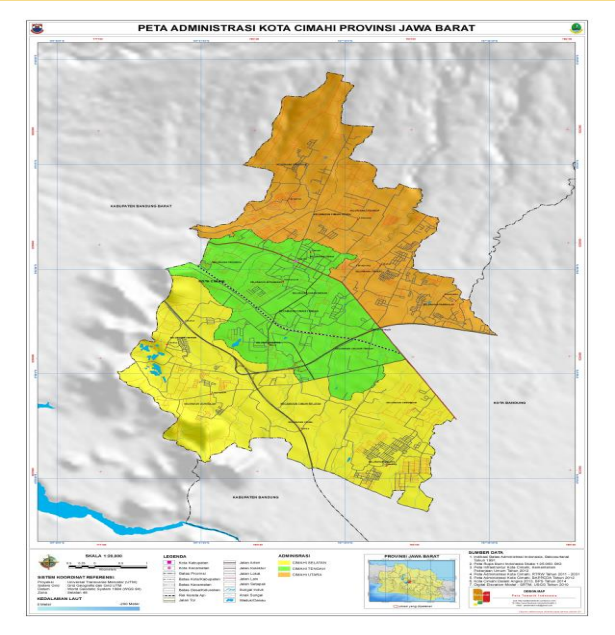
Background

Urban traffic has always been a main challenge for highly populated city like Cimahi City, West Java, Indonesia with its near 600.000 inhabitants (2021).

With the increase of traffic congestion, it will directly cause many complications from pollution, longer journey duration, and health risk.

Here we want to help urban planner of Cimahi City in planning better traffic flow to ease the current roadways with heavy traffic jams.

We try to do this by using algorithm that can cluster main roadways by its congestion trend and severity, therefore we can use the result to optimize road routes and logistics



Objectives

Business Objectives

To help urban planner of Cimahi City in mapping Cimahi's important street by their traffic flow congestion trend during traffic jams.

Model Objectives

Creating clustering model to group important/frequently jammed streets at Cimahi City based on their traffic flow congestion trend.

Success Criteria

Model manages to identify well-clustered group of streets based on their traffic flow congestion trend.

Data Understanding

A. Dataset Used

Aggregated jams Waze for Cities data `aggregate_median_jams` for Cimahi City from 2022-07-06 00:00:00 to 2022-09-06 00:00:00 which contains median traffic speed, median delay time, and median jam length of reported traffic jams.

B. Calculated Fields Definition

- **hour :** Hour of traffic
- **day_type :** Day of the week classification (Weekday or Weekend)
- **traffic_flow** : Reciprocated aggregated traffic speed during traffic jams
- **label** : cluster group label

C. Data Processing Flow

1. Data Preparation

- Filtering
- Feature engineering
- EDA
- Data Transformation
- Data Aggregation

2. Data Cleaning

- Cardinality Reduction
- Imputation
- Scaling

3. Data Modelling

- kMeans without Dimensionality Reduction
- kMeans with Dimensionality Reduction

D. EDA

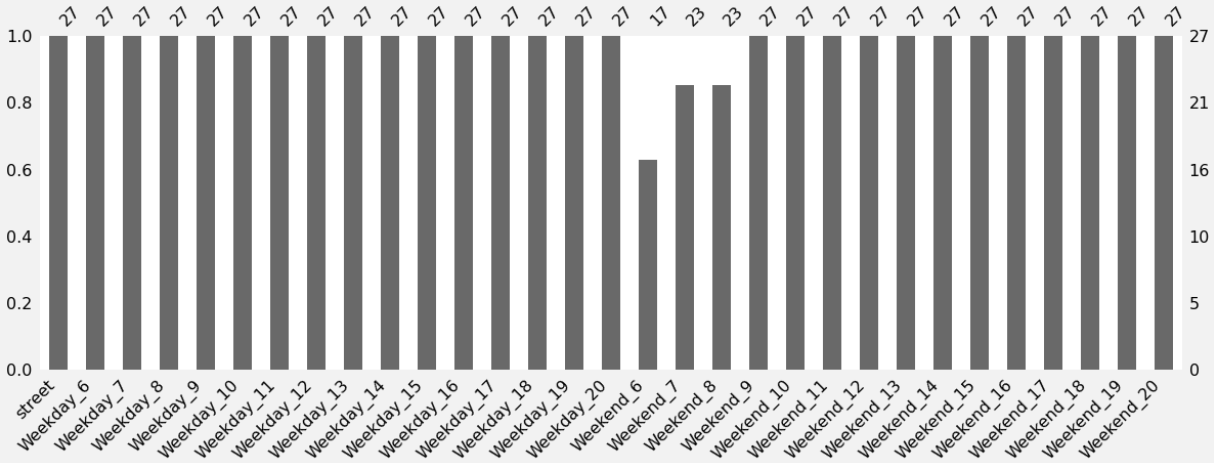
Insight	Impact
Traffic flow trend can't be grouped by day (Sun-Mon) due to sparsity of data, even for street with high amount of records	Trend by day categorized as day type. - Mon-Fri → Weekday - Sat-Sun → Weekend
Number of records distribution for each street are highly skewed.	Select only "important" street name (n_records >1% total_records)
Batu Basal street always has 0 for the records of traffic jam speed	Always in blockade during the timeframe, can be removed.
To be able to fill missing aggregated values, traffic speed needs to be reciprocated so that higher traffic speed lies closer to 0	Dimensions use for training are in their reciprocated values

E. Data Cleaning

Steps

Reducing street cardinality by selecting only streets with >1% total records (important streets)

Imputing missing value with 0



Robust scaling reciprocated traffic jams speed

Impact

Number of streets : 195 → 27

No missing value, important for kMeans algorithm

Better cluster quality, treat outliers

Modelling

1st Approach : kMeans without Dimensionality Reduction

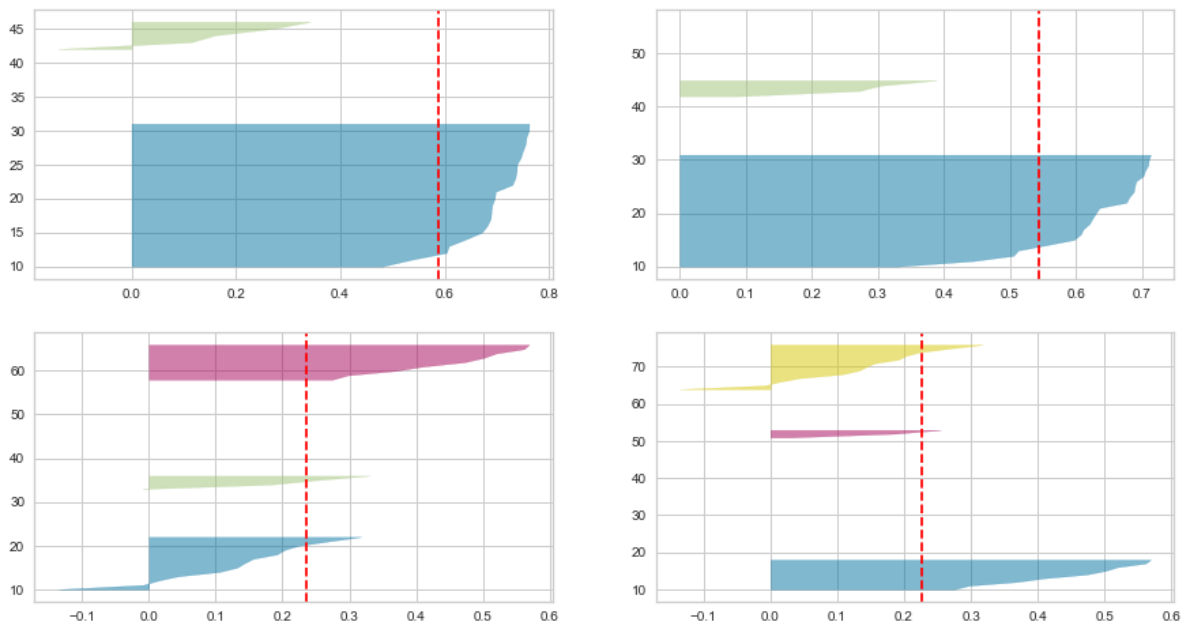
Steps :

1. Optimize number of clusters (k) with elbow method & silhouette score
2. Evaluate clustering performance

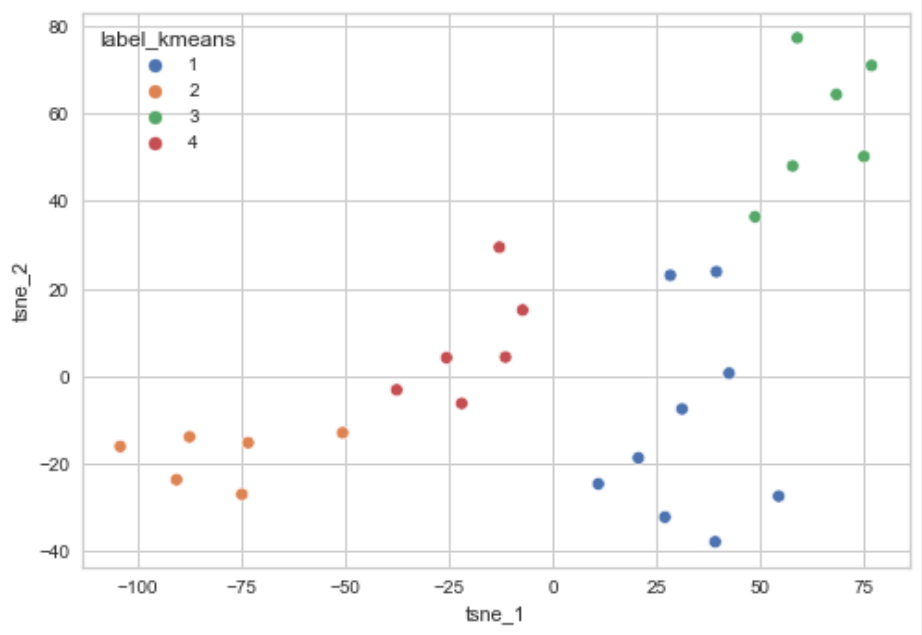
Results :

All inspected k (2-7) resulted in suboptimal separation for given data due to :

1. Presence of clusters with below average silhouette scores
2. Wide fluctuation in silhouette size
3. Occurrence of negative score



2nd Approach : kMeans with Dimensionality Reduction



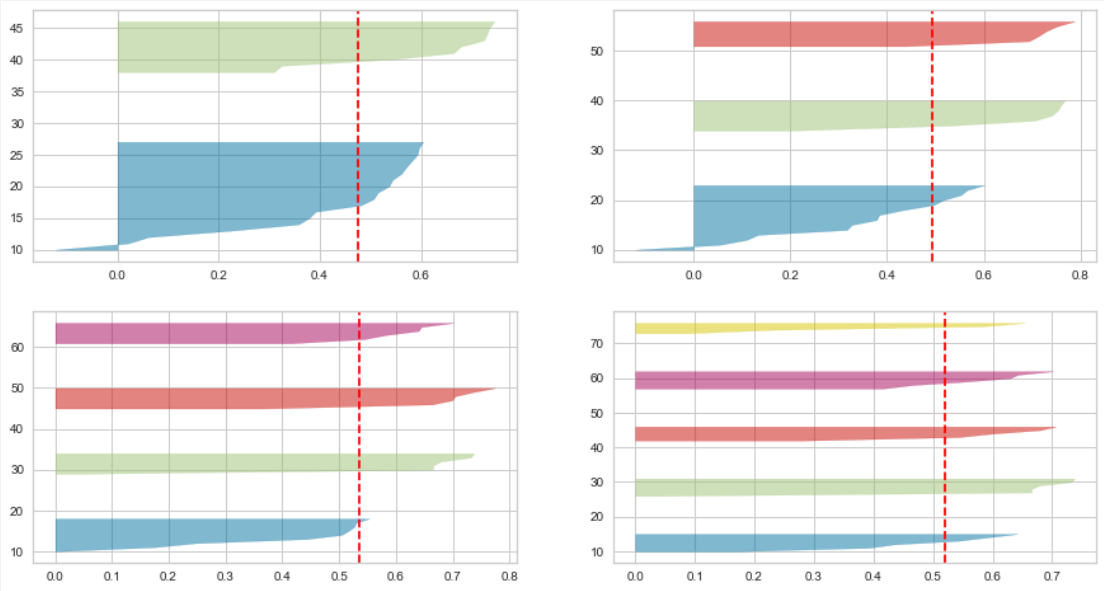
Steps :

1. Apply tSNE method to reduce data dimensions from 30 to 2
2. Optimize number of clusters (k) with elbow method & silhouette score
3. Evaluate clustering performance

Results :

Good performance for k=4.

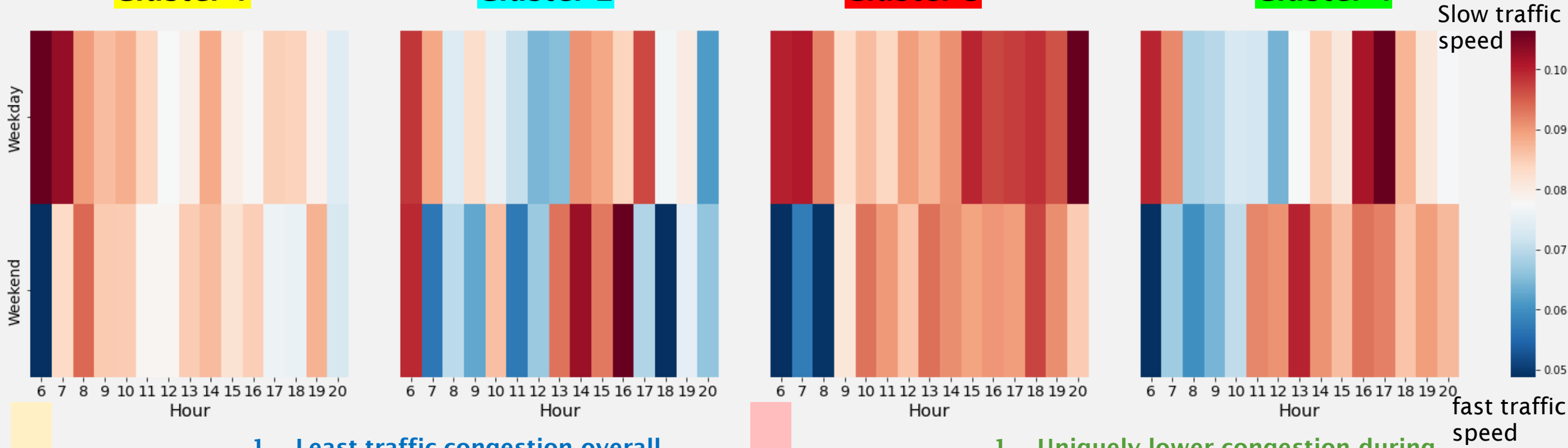
Evaluation metrics :
- Inertia : 10682.330078125
- Silhouette Score : 0.5360958



Results

Traffic Flow Trend by Clusters

Cluster 1 **Cluster 2** **Cluster 3** **Cluster 4**



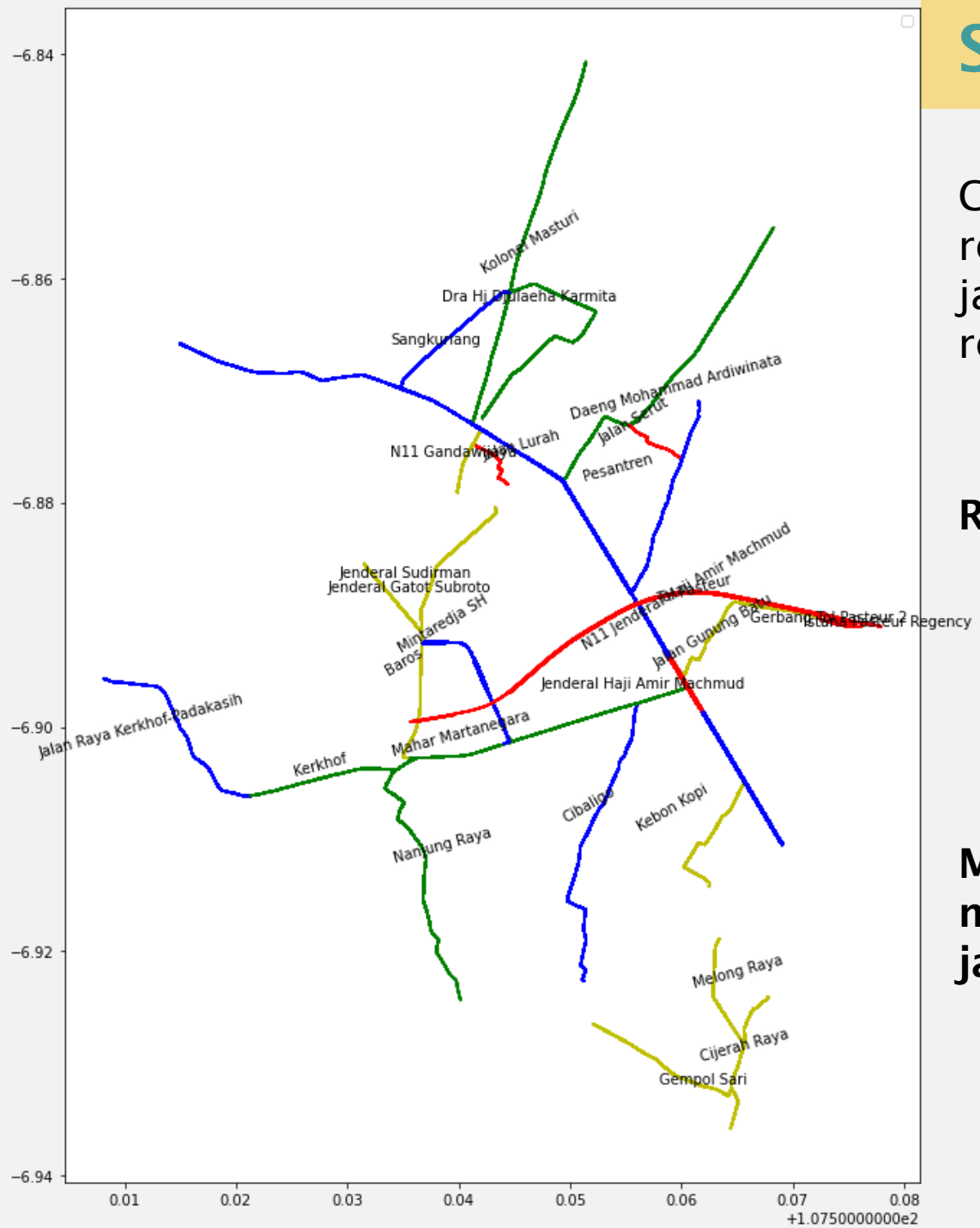
1. **Least traffic congestion overall** (worse congestion during rush hour).
2. worst congestion during Weekend evening

1. **Uniquely lower congestion during weekday.**
2. On Weekday only started having worse congestion after noon.

1. **Overall bad traffic congestion both Weekday and Weekend**
2. Very bad congestion on Weekday morning

1. **Worst traffic congestion overall** (only low during weekend morning)
2. During Weekday average jams speed is around 5-10 Km/h

Street Map by Cluster



Cimahi roadway in general has healthy amount of roads that are not too congested during traffic jams, but there are some keynotes on important roads with severe congestion during traffic jams.

Roadways with the worst average traffic jams :

- Tol Pasteur – N11 Jenderal Haji Amir Machmud
- Jalan Serut
- Jalan Gunung Batu
- Jalan Lurah

Main roads on the southern parts of Cimahi mostly have moderate congestion during traffic jams.

Summary

1. We've manage to **cluster important streets in Cimahi City** by its average traffic flow trend using kMeans algorithm + tSNE dimensionality reduction method.
2. **4 Clusters was formed**, with cluster 3(red) and cluster 1(yellow) indicating streets with high intensity of traffic jams and bad congestion during traffic jams.
3. **Cluster 4 has unique properties** where its streets usually has less severe traffic jams during Weekday than Weekend.
4. **Roadways with the worst congestion** are Tol Pasteur – N11 Jenderal Haji Amir Machmud street and other short street north of the city
5. **Main roadways in the southern part** of the city mostly have moderate congestion during traffic jams.
6. **Further exploratio**n can be done by connecting the clusters to other metrics like delay, weather, alerts, and irregularities. Also how well congestion speed distribution differs from one cluster to another can be tested using inferential statistics

Reference

https://opendata.jabarprov.go.id/id/artikel/menilik-prediksi-arus-mudik-idulfitri-2022-di-jawa-barat

https://github.com/MikeS-nth/portfolio/tree/master/DC_Traffic_Prediction

https://pennmusa.github.io/MUSA_801.io/project_8/index.html#2_data

https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556#:~:text=DBSCAN%20stands%20for%20density%2Dbased,many%20points%20from%20that%20cluster.

https://towardsdatascience.com/lets-do-spatial-clustering-with-dbscan-c3dbfd9fc4d2

https://towardsdatascience.com/k-means-algorithm-for-high-dimensional-data-clustering-714c6980daa9

https://towardsdatascience.com/how-to-tune-hyperparameters-of-tsne-7c0596a18868

https://iopscience.iop.org/article/10.1088/1755-1315/31/1/012012/pdf

How Data Led Urban Planning Is Changing the Urban Landscape (otonomo.io)

Microsoft Word - 9389 Tamplate- (semanticscholar.org)